Signal-Detection Studies, with Applications

By E. L. KAPLAN

(Manuscript received October 10, 1954)

Curves are given relating the probability β of detection of a signal in noise to the signal-to-noise power-ratio S/N, to the proportion α of false detections that can be tolerated, and to the time available (more specifically, to the number m of independent squared samples of the envelope of the filter output that are averaged in making a single attempt at detection). For the curves, three types of sinusoidal signals are assumed, according as the amplitude is constant, varies at random very slowly (fading), or varies as rapidly as the filtered noise itself. The second case is of course very unfavorable for reliable detection, and the third is also if one is limited to one or two sample points. The curves are applied to problems of optimizing radar parameters such as pulse energy, scan rate, averaging time, etc.

The Appendix gives the mathematical background for the foregoing and then proceeds to consider additional types of signal (dc and arbitrary Gaussian) and methods of detection (failure to rectify or take the envelope, counting of samples above a threshold, and averaging by continuous integration).

1. INTRODUCTION

The word detection suggests a decision between just two alternatives: either a signal was transmitted, or it was not. Typically, however, such decisions are made repeatedly at short intervals of time, with the effect that a multiplicity of possibilities are involved, namely the time intervals in which signals are received, and perhaps also the carrier frequencies at which they are transmitted. In radar and sonar systems, rotating directional receivers map one, two, or three dimensions of space upon the time axis, so that the observation of time is translatable into an observation of position. In these applications, of course, the received signal is a reflection of one transmitted by the observer, and the aircraft or other object to be detected is not attempting to communicate with the ob-

server. Telegraphy and PCM are applications in which the signals do represent attempts to communicate. In all cases, random noise is present, and will hide signals whose intensity is sufficiently low.

If the bandwidth of the noise accompanying the signal is considerably larger than that of the signal, the best practical procedure is well known to be to pass the signal-plus-noise through a bandpass filter whose bandwidth approximates the larger of the following: (a) The bandwidth of the signal. (b) The reciprocal of the time available for detecting the signal.[1] If the frequency of the signal is not known in advance, one of course moves the pass-band of the filter up and down the frequency scale and thus searches for the signal in frequency as well as in time; the present discussion applies equally well to this case.

In this way much of the noise may be eliminated, but that is not the end of the problem. Some noise still gets through. However, the presence of a signal may be expected to produce an increase in the magnitude of the rectified output. The purpose of this discussion is to indicate how great this increase should be required to be before one decides a signal is present, to draw conclusions therefrom regarding optimum procedures, and to illustrate the great importance of the nature of the signal and the method of detection. For simplicity, square-law rectification is assumed; it is known that linear rectification does not give very different results.[2] The noise, and in some cases the signal also, is assumed to be Gaussian.

We assume the following detection procedure, which may apply literally or else be approximately equivalent to that used in a physical system (e.g., averaging by continuous integration, as in Section 15; or by a human observer, or a phosphor). The rectified output is sampled at m different instants of time,[3] which are far enough apart so that the values of the noise are effectively independent (uncorrelated). If the average output at these m instants exceeds the average noise level N by an amount kN or more, we decide that a signal is present; if not, we decide no signal is present. For simplicity, the value of N is assumed to be known as the result of past observation; since this is only approximately true in practice, the results will somewhat exaggerate the effectiveness of the detection procedure, especially for large values of m and low signal strengths. Another assumption made in calculating the curves is that if

[1] Alternative (b) is superfluous if, in determining the bandwidth, the signal is defined to be zero outside the time interval within which it is available for analysis.

[2] M. Schwartz in his Harvard (1951) dissertation, "A Statistical Approach to the Automatic Search Problem," finds agreement to within 0.2 db.

[3] This assumes a narrow-band output whose envelope is sampled; if no envelope is obtained, the number of samples is denoted by $2m$. The point will be discussed below.

a signal is present at all, it is present at each of the m instants sampled; the methods for removing this assumption are discussed later.

It is well known (and shown in Section 5) that the averaging after rectification described above does not fully compensate for a filter whose passband is too wide; but when the filter has been made as selective as it can or should be, the post-rectification averaging, when it is possible, gives a further increase in sensitivity.

In practice many of these averages must be formed, and many decisions made, corresponding to the search for the signal in time and possibly also in frequency and in geometrical position (e.g., radar detection of an airplane). The probability (for any one average or any one decision) of deciding that a signal is present is denoted by β or α according as the signal is actually present or not. Thus α is the proportion of false alarms among the total number of averages containing no signal, while β is the proportion of valid detections among all those averages that do contain the signal. A detection is considered valid only if it is associated with the average, or one of the averages, in which the signal actually occurs. The average time between decisions divided by α gives the average time between false alarms (in case the signal is usually absent).

The problem could now be formulated thus: to calculate the values of k and β corresponding to given values of m, α, and the signal-to-noise power-ratio S/N. Actually it has been convenient to regard m and S/N as the principal variables, and so to plot their relationship for a few different values of β and α. The number k depends only on m and α, and is numerically equal to the S/N values (*not* measured in db) given by the curves for $\beta = 0.50$ in Fig. 1. (These curves should be rigorously correct for k, whereas their use for S/N involves an approximation as discussed below.)

2. THE THREE TYPES OF SIGNAL AND EXPLANATION OF THE GRAPHS

Three kinds of signal are considered: (1) Steady sinusoid. (2) Fading sinusoid. (3) Noise-like signal. All of these are in a sense special cases, but other signals will generally be of some type intermediate to these.

Cases 1 and 2 are alike in that for both it is assumed that the m signal amplitudes occurring in an average are identical, whereas in Case 3 these m amplitudes are assumed to be independent random variables, just as the corresponding noise values are.

Cases 1 and 2 differ in that in Case 1 the signal-to-noise ratio S/N refers to the instantaneous signal power received, while in Case 2, S/N

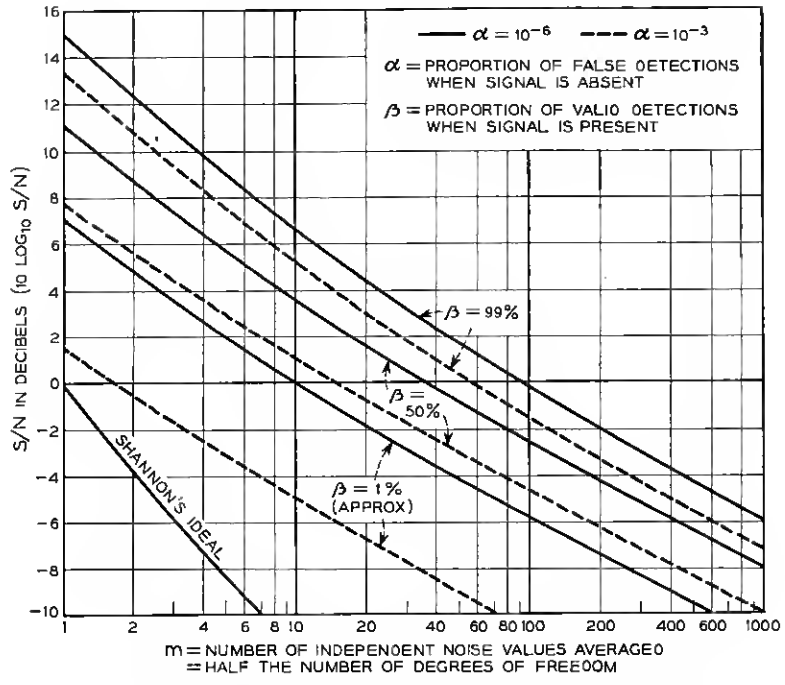

Fig. 1 — Signal-to-noise ratio required for detection of steady sinusoidal signal.

refers to a long-term average of the signal power, which may be either greater or less than the instantaneous power. Thus the signal amplitude in Case 2 is a random variable, but one that does not change much during the period of time that is averaged over. The Rayleigh or circular Gaussian distribution has been assumed for the signal amplitude, corresponding to a Gaussian signal. For the signal power, this becomes the exponential distribution.

The pulses of carrier used in communication, and possibly the radar return from a fixed object, can be considered as steady sinusoids (Case 1). The radar return from an airplane belongs to Cases 2 or 3 or intermediate cases, since the phase relations of the returns from different parts of the airplane change by different amounts as the aspect of the airplane changes. Case 2 may arise in at least three ways:

(a) The signal as generated may be Gaussian and have a very narrow but non-zero bandwidth; i.e., its Nyquist interval may be as large as, or larger than, the averaging time employed.

(b) The signal may consist of several sinusoids of slightly different

frequencies beating against one another. If there are reflecting surfaces involved and the source of the signal, the observer, and the reflecting surfaces have any relative motion, the same phenomenon of beats occurs even with a single sinusoid, whenever multiple paths for its transmission exist. Interference between radar reflections from an airplane and from its image in the sea surface is a familiar example.

(c) Even if the signal in each instance is a steady sinusoid, one may still have to regard the collection of amplitudes encountered on different occasions as values of a random variable. Thus, since various airplanes have different speeds and radar cross-sections and are detected at different ranges, the strengths of their radar returns will differ accordingly. Another example is provided by the interference between direct and reflected signals in the absence of any relative motion among the signal

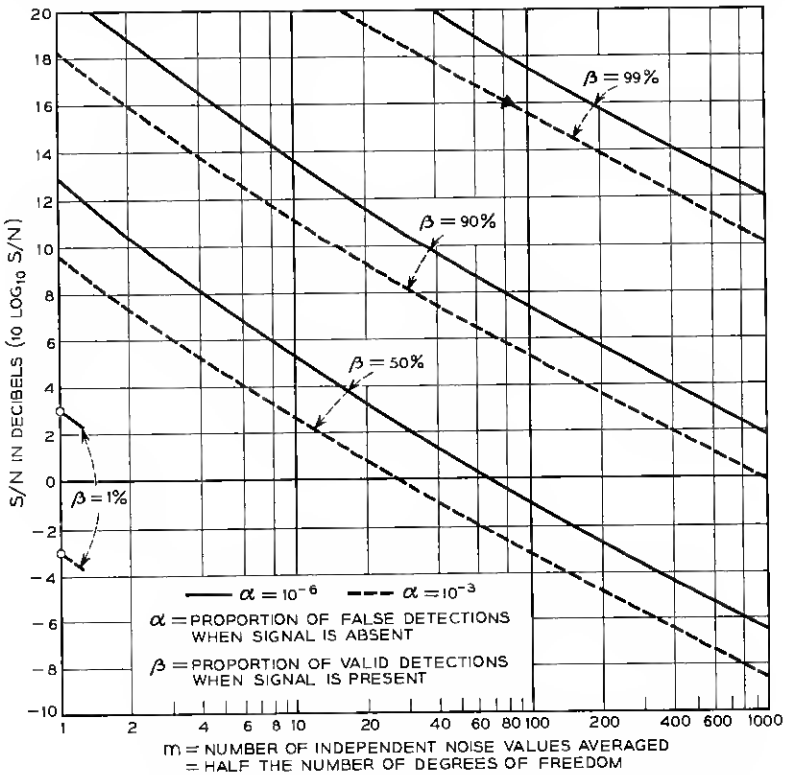

Fig. 2 — Signal-to-noise ratio required for detection of a fading sinusoidal signal (amplitude steady during each detection, but variable from one detection to another).

source, the observer, and the reflecting surface. Then the signal strength may be constant with time, but it will still be a random variable depending on the positions that the source and the observer happen to occupy.

The exponential distribution of signal power assumed in the present treatment of Case 2 is theoretically exact for situation (a) above, and possibly as good as one can do for situation (b). In situation (c) the proper distribution may be quite different, though its effect should be qualitatively similar.

In all cases the S/N value is defined as if the signal were continuously present; i.e., it does not depend on tbe duration of the signal. In Case 1, S is the peak signal power, while in Case 2 it is the average of the various possible peak signal powers. In Case 3, S might be described as the peak value of the statistical expectation of tbe power.

Curves of S/N versus m are given for the three cases in Figs. 1 to 3

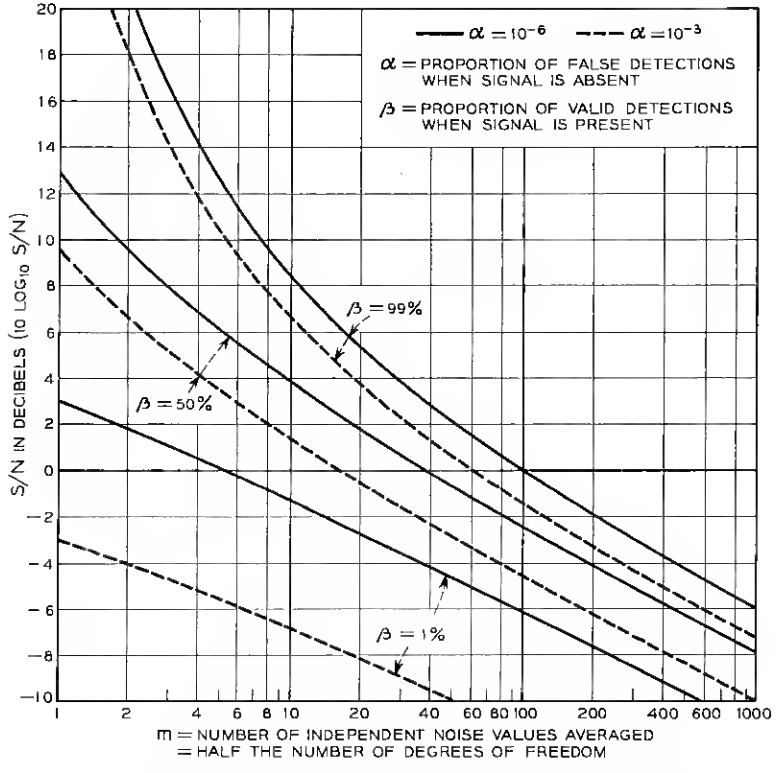

Fig. 3 — Signal-to-noise ratio required for detection of a noise-like signal.

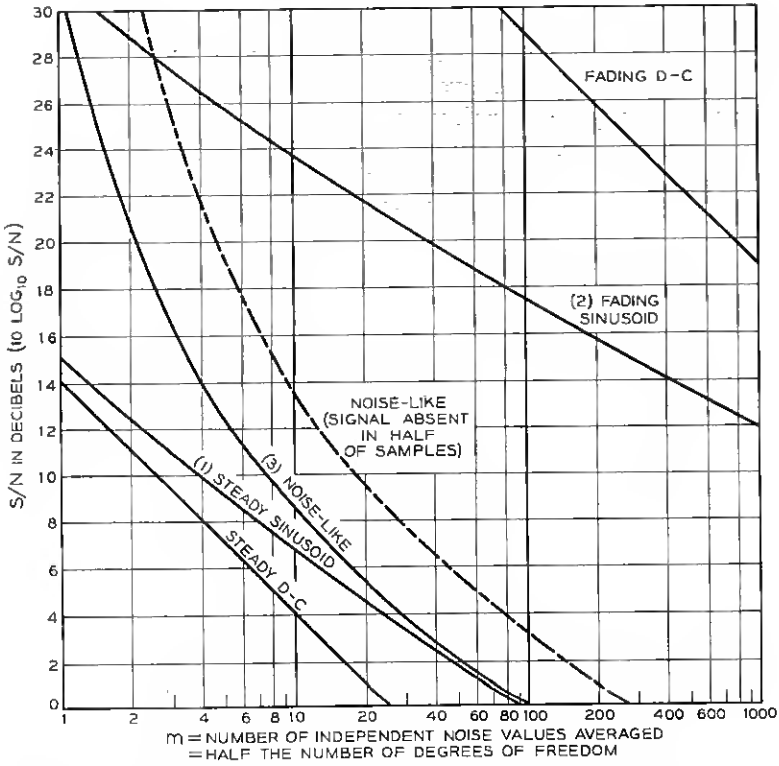

Fig. 4 — Comparison of the signal-to-noise ratios required for detecting various kinds of signals.

respectively, for $\alpha = 10^{-3}$ (dotted lines) and 10^{-6} (solid lines) and three values of β. In Fig. 4 one curve from each of the other figures is reproduced for comparison, and also one new curve (the broken line) to be discussed in Section 6. Curves for $\beta = 90$ per cent may be interpolated in Figure 1 (and in Fig. 3 when $m \geqq 10$) at about 0.58 of the distance from the curve for $\beta = 50$ per cent toward the curve for $\beta = 99$ per cent. This may be done by eye, or numerically with S/N expressed in decibels. The curves for $\beta = 1$ per cent on Figure 1 are probably of low accuracy.

These same three cases are also considered (for somewhat different values of the parameters) by M. Schwartz in his dissertation cited previously, and by J. I. Marcum and P. Swerling in their work at the Rand Corporation.

The curve in the lower left corner of Fig. 1 shows the relation

$$S/N = 2^{1/m} - 1$$

that would exist between S/N and m if the maximum information-carrying capacity of the system were utilized. The maximum capacity is $(1/2m)$ bit per Nyquist interval ($1/2W$; see next Section), achieved by having the signal present with an independent probability of 1/2 in each average of m samples. According to C. E. Shannon's theorem,[4] the maximum capacity in terms of S/N is $\log_2(1 + S/N)^{1/2}$ bits per Nyquist interval. When m is large (or equivalently, S/N small), the relation is $S/N \doteq (\log_e 2)/m$.

3. DETAILS: DC SIGNALS, ENVELOPES, ETC.

This section discusses various other cases of less importance than the three defined in the preceding section; namely, dc signals, low-pass and broadband filters, the relations between instantaneous values and the envelope, and the distinction between narrow-band noise and a noise-modulated carrier. The first three topics are intimately connected; just as sinusoidal (ac) signals, narrow-band filters, and envelopes naturally go together (Sections 1 and 2), so are dc signals, low-pass filters, and instantaneous samples usually associated.

The optimum detection of a dc signal, given independent samples, is trivially easy. No filter is needed, and rectification is undesirable, as shown by the two lowest curves on Fig. 4. The lowest curve, labelled "steady dc," is obtained by the optimum procedure of taking the simple average of the unrectified sample values, some of which may be negative. If the squared values are averaged, the curve will coincide with the next higher curve, labelled "steady sinusoid," as shown later in connection with equations (13b) and (13c). The "fading dc" signal is one which is constant during each detection but has a Gaussian distribution of mean zero from detection to detection. Although it is the worst curve on the figure, quadratic averaging would make it still poorer. [On the other hand, (12c) shows that rectification is as indispensable for the detection of a lowpass noise-like signal as for any other rapidly oscillating signal.] Both of the dc curves are straight lines of 45° inclination, and for them the number of samples is not m but $2m$. This adjustment keeps the total time the same, and in the case of quadratic averaging produces complete equivalence between the dc and the sinusoidal signal, as shown in Section 13. The simple formulas applying to the averaging of unrectified dc signals are given later as equations (12a) to (12c).

If a filter is used in detecting a dc signal, it must be of the low-pass

[4] B.S.T.J., **27**, pp. 379–423, 623–656, 1948. Proc. Inst. Radio Eng. **37**, pp. 10–21, 1949. Mathematical Theory of Communication (with Warren Weaver). Univ. Illinois Press, 1949.

type,[5] and its effect is not essentially different from that of the assumed averaging of m sample values. In detecting an ac signal, a narrow-band filter is decidedly preferable to a low pass because it eliminates more noise; but it also has the effect of permitting the (squared) envelope of the filter output to be obtained, provided the frequencies contained in this output are confined to a range of less than 3 to 1. In this case the output of the square-law rectifier will contain two separate bands of frequencies of equal widths which (given a reasonable separation between them) can be separated by another filtration. The lower frequency-band is usually the only one that is wanted, and it gives half the square of the envelope of the output of the first filter. If the frequency range of the first filter output is 3 to 1 or more, the envelope is still susceptible of a theoretical definition but it cannot be isolated so easily and is then a somewhat artificial concept.

The general effect of a second filtration after rectification of an ac signal, whether or not the envelope is actually obtained, is to increase the reliability of a single sample, without necessarily having a marked effect on an average over a long (fixed) period of time. The elimination of the sinusoidal component resulting from the signal, and having double its original frequency, is perhaps the main object. If this is not done, it will be necessary to do considerable averaging to eliminate the possibility of missing the signal because of sampling it at its troughs. In this respect instantaneous sampling of a sinusoidal signal is qualitatively similar to the various cases of the sampling of the noise-like signal.

The properties of the envelope will now be considered. Let $n_1(t)$ and $n_2(t)$ be Gaussian noises containing no radian frequency above $\omega/2$. Then the spectrum of

$$n(t) = n_1(t) \cos \Omega t + n_2(t) \sin \Omega t \tag{3a}$$

is limited to the interval $\Omega - \omega/2$, $\Omega + \omega/2$. Conversely, consideration of the Fourier transform of $n(t)$ shows[6] that any (stationary) noise whose spectrum is limited to $\Omega \pm \omega/2$ can be expressed in the form (3a). It is found furthermore that

$$En_1(t)n_1(u) = En_2(t)n_2(u) = \int_0^\infty P(\lambda) \cos (\lambda - \Omega)(t - u)\, d\lambda \tag{3b}$$

$$En_1(t)n_2(u) = -En_1(u)n_2(t) = \int_0^\infty P(\lambda) \sin (\lambda - \Omega)(t - u)\, d\lambda$$

[5] This may not be true in a rigorous sense. If indefinitely prolonged dc signals are not encountered, a cut-off at a sufficiently low frequency is permissible.

[6] For example, see S. O. Rice, B.S.T.J., **24**, pp. 46–156, 1945, Section 3.7.

where $P(\lambda)$ is the power per radian/sec. for frequencies in $n(t)$ whose absolute values are in the neighborhood of λ radians/sec. Let $\psi_0(t - u)$ and $\psi_1(t - u)$ represent the two distinct covariances of (3b), the subscripts 0 and 1 reflecting the even and the odd characters of the two functions. Then

$$En(t)n(u) = \psi_0(t - u) \cos \Omega(t - u) - \psi_1(t - u) \sin \Omega(t - u) \quad (3c)$$

Replacing Ω in (3a) by $\Omega_0 + \Omega_1$ gives

$$n(t) = [n_1(t) \cos \Omega_1 t + n_2(t) \sin \Omega_1 t] \cos \Omega_0 t + [-n_1(t) \sin \Omega_1 t + n_2(t) \cos \Omega_1 t] \sin \Omega_0 t$$

so that the same envelope

$$\sqrt{n_1^2(t) + n_2^2(t)}$$

is obtained regardless of the value adopted for the "carrier" frequency. However, the highest frequency occurring in $n_1(t)$ and $n_2(t)$ will be minimized by taking Ω in the center of band; and (3b) shows that if the power spectrum of $n(t)$ has a center of symmetry which is taken as Ω, then $En_1(t)n_2(u) = 0$.

The noise-like signal used in this study, like the filtered noise itself, has the form (3a). The curves do not apply quantitatively to a single term $n_1(t) \cos \Omega t$ obtained by modulating the carrier $\cos \Omega t$ with a low-frequency Gaussian noise-like signal $n_1(t)$, though the necessary tools are given in equations (13g) and (13h) and the remarks following thereafter.

The most obvious difference between the two signals resides in their envelopes, which are

$$\sqrt{n_1^2(t) + n_2^2(t)}$$

(with a Rayleigh distribution) and $|\, n_1(t) \,|$ (the absolute value of a Gaussian variable) respectively, but other differences may be discovered. The former can be, and is assumed to be, stationary, Gaussian, and ergodic. The latter is stationary only if the phase φ of the carrier $\cos (\Omega t + \varphi)$ is taken to be random, but the relative phases of the sinusoidal components of $n_1(t) \cos (\Omega t + \varphi)$ are still not completely random, and the signal is neither Gaussian nor ergodic. [The Gaussian variable $n_1(t)$ is multipled by $\cos (\Omega t + \varphi)$, which has the distribution $dz/\pi\sqrt{1 - z^2}$.] Where detection is concerned, the two cases are qualitatively similar, but the single term $n_1(t) \cos \Omega t$ has a greater amount of fluctuation for the same amount of power, and is therefore a little harder to detect than the sum of the two terms.

The results of this study have been phrased in terms of the narrow-band two-term form $n_1(t) \cos \Omega t + n_2(t) \sin \Omega t$ for the noise and the noise-like signal. However, the curves apply also to the less important case of the detection of a dc signal by rectification (which is not the optimal procedure, as shown above). The word "sinusoid" is to be replaced by "dc signal," and *two* independent samples of the rectifier output are to be taken for every one that is prescribed in the narrow-band case. From a noise or signal that (at the input to the square-law rectifier) has a flat spectrum of width W cps, independent samples are obtainable at interval $1/W$ seconds in the narrow-band case, and at interval $1/2W$ in the lowpass (dc) case,[7] so that for large m the averaging time is m/W seconds in both cases. The amplitude of the "fading" signal of Case 2 with its Rayleigh distribution is equivalent in the lowpass case to the RMS value formed from two independent samples of a Gaussian noise-like signal. The case of only one sample could be calculated but would give even less satisfactory detection.

4. NUMERICAL EXAMPLES (RADAR)

The following hypothetical example illustrates the concepts involved. Suppose one has a radar that searches in range and azimuth with a beam 3° wide, a pulse rate of 5,000 per second, a pulse length of 1 microsecond, and a scan rate of 10 per minute. On the average one false detection in 15 minutes can be tolerated. Pulses returned by a target during a single scan are averaged (after rectification). What signal-to-noise ratio is required to give a probability of 90 per cent of detection in one scan?

This is the narrow-band case. The pulse length is usually about equal to the reciprocal of the bandwidth and so can be taken as the sampling interval. 200 samples of the radar return could then be taken in the 0.0002 sec. elapsing between consecutive transmitted pulses. Suppose that 150 of these are relevant, giving 150 values of range that can be distinguished. Similarly the azimuth scan of 360° is covered by 120 beam-widths (this does not assume that the azimuth accuracy is no better than 3°). If a factor of 2 is allowed for overlap (in range or azimuth or both) among the averages, the number of decisions per scan is $2 \times 120 \times 150 = 36{,}000$, or 5,400,000 in a period of 15 minutes. So $\alpha = 1/5{,}400{,}000 = 2 \times 10^{-7}$. In each scan the beam is on a target for 0.05 sec., the time required to turn 3°. In this time 250 pulses are transmitted; this is the value of m,

[7] The latter interval is well-known; the former is twice as long because the smoothing accompanying rectification in the narrow-band case makes the sampling equivalent to the sampling of $n_1(t)$ and $n_2(t)$, and their bandwidth is not W but only $W/2$.

since returns from different ranges and/or azimuths are averaged separately (e.g., by a PPI scope). The 250 samples in an average are thus not consecutive in time in this case.

The values of S/N can be read just above the curves for $\alpha = 10^{-6}$ (which are the solid curves). For a non-fading return, Fig. 3 then gives S/N equal to -4.2 db for $\beta = 50$ per cent and -2.2 db for $\beta = 99$ per cent; by interpolation, S/N is about -3 db for $\beta = 90$ per cent. It should be remembered that S is the peak (not the average) signal power. For a fading return, Fig. 2 gives S/N equal to 5.5 db. Use of the latter figure would imply that the returned pulses had virtually the same amplitude during the time of 0.05 sec. during each scan when the beam was on the target (but still had a random amplitude when viewed for a sufficiently longer period). On the other hand, use of S/N equal to -3 db would imply that the values of the signal at interval 0.0002 sec. were virtually independent samples. These conditions may also be expressed in terms of the width of the lines or narrow bands, if any, in the spectrum of the signal (the spacing between the lines is the pulse repetition rate of 5000 cps, which is irrelevant here). If the width is 5000 cps or more (i.e., the spectrum continuous), one has S/N equal to -3 db; if the width is 20 cps or less, one has S/N equal to 5.5 db. The intermediate region is rather extensive, corresponding to a factor of $m = 250$. It might be explored in an approximate manner by the formulas (13g) and (13k).

If the line-width is not much more than 20 cps, but not less than say 1 cps, then conclusion (A) of Section 8 shows that the probability of detecting the airplane within 6 seconds ($=1$ scan, as first stated) is increased by increasing the scan rate to as much as 1 per second, so that a number of scans are completed within the 6 seconds.

Suppose now that one has a radar receiver with an antenna pattern 3° wide scanning in azimuth at 10 rps as before, whose object is to detect a distant radar C-W transmitter whose frequency is only known to lie within a band 150-mc wide. The detection is accomplished by passing the signal through a filtering device that passes 150 different frequency bands of 1 mc each in succession (one at a time). The problem is then numerically the same as before, the search in frequency having replaced the search in range. There is the difference that here the samples in any average are presumably all consecutive in time, while previously they were taken at intervals of 0.0002 sec., the time between transmitted pulses. The search in frequency also occurs in the determination of the velocity of an airplane by its Doppler frequency, but the bandwidth of the frequency-analyzer is then much less than 1 megacycle, and not nearly so many independent samples can be obtained.

If the signal is applied simultaneously to a bank of filters (e.g., vibrating reeds), the number of decisions is the same as before, but the saving of time would permit a corresponding increase in the number m of samples in an average.

B. McMillan has pointed out that with long-range search radars whose resolution is much better in range than in the other coordinates, some of the range resolution might profitably be sacrificed in order to increase the pulse length. Although the longer pulse length could be used to increase the value of m, it would be preferable to decrease the receiver bandwidth instead, and thus decrease the noise power N in the same ratio.

Conclusions

5. THREE MAJOR EFFECTS

The curves show that the steady sinusoid is the easiest signal to detect, as one would expect, while the fading sinusoid is the most difficult to detect reliably. It is therefore very desirable to avoid the latter situation if reliable detection is wanted, either by reducing the severity of the fading and so moving toward Case 1, or by sampling the output over a period of time long enough to average out some of the fluctuation in the signal amplitude, and so moving toward Case 3 (noise-like signal). The latter is of intermediate difficulty of detection, and coincides with Case 2 when $m = 1$ and approaches coincidence with Case 1 as $m \to \infty$. The difficulty with the fading sinusoid is this: If the average signal power is equal to that which gives good detection in Case 1, little reliability is left to be gained in those detections where additional signal power happens to be available, but much may be lost when the signal power happens to be lower than the average.

A second major conclusion results from the steepness of the left-hand portion of the curves for $\beta = 99$ per cent in Fig. 3: If reliable detection of a noise-like signal is required, it is highly desirable that at least 4 or 5 *independent* samples of the signal be available and made use of. In fact, if a 99 per cent chance of detection is desired, 4 samples require only $\frac{1}{50}$ of the signal power, or $\frac{1}{12}$ of the energy required by one sample. The principle is well-known in connection with search radars, which are designed to return about four pulses from a target.

When m is significantly greater than unity, and the samples being averaged are adjacent to one another in time, one has the option of using a more selective filter and thus doing the averaging linearly (with preservation of signs) before rectification rather than after. The former is well-

known to be the more effective in Cases 1 and 2. (Case 3, on the contrary, represents the situation that results when the selectivity of the filter has already been made as great as it can profitably be.) Let the time available for the detection be fixed. Then m, the number of independent noise samples available, is proportional to the filter bandwidth; the latter is inversely proportional to S/N, provided the noise spectrum can be regarded as flat in the region considered, and the filter does not reject any of the signal frequencies. Thus m and S/N are inversely proportional, and one is operating along a 45° line in Figures 1 and 2 (both m and S/N being plotted logarithmically). Evidently β is maximized by keeping m (and hence the bandwidth) small.

6. MATCHING THE DURATIONS OF SAMPLE AND SIGNAL

It is fairly obvious intuitively that the chance of detection would be greatest if some one average coincided exactly with the period during which the signal was present. (The signal may be intermittent, as with the first radar of Section 4; if so, the sampling also should be intermittent in the same pattern.) There is no way to insure this if the duration of the signal is not known in advance, but if it is known, the desired coincidence could be approximated to some extent by having the averages overlap; e.g., run from 0 to 1 time unit, ½ to 1½, 1 to 2, 1½ to 2½, etc. This increase in n necessitates use of a smaller value of α and hence requires an increase in signal strength, but the latter increase is insignificant: When $\beta \geqq 0.50$, the curves show that it is 0.3 db or less for a factor of 2 in α. Similarly, the passbands may overlap when one searches in frequency, and the radar antenna patterns may overlap when one searches in direction.

If the averages include fewer samples (though all the samples in some average contain the signal), the effect in any of the Figs. 1 to 4 is to move to the left along a horizontal line. The detection probability β is decreased because there is no increase in the concentration of the signal energy, while the fluctuation of the noise is averaged out less effectively.

If the averages include too many samples, then some of these samples will always consist of pure noise, even when a signal is present, and the concentration of the signal energy is reduced relative to that of the noise. If a number m' of samples containing the signal are increased to m by adding $m-m'$ samples of pure noise, the effect in Figs. 1 and 2 is to move down and to the right along a 45° line, since the signal-to-noise ratio is effectively reduced by the same factor m/m' by which the number of samples is increased. Evidently this decreases β.

For the noise-like signal of Fig. 3, new calculations are required when $m' < m$, and the broken line in Fig. 4 shows some results obtained by the chi-square approximation of equation (13g), with $m'' = m'$. This curve is plotted against m (assumed for this example to be twice m') and so shows the effect of halving the duration of the signal. To show the effect of adding an equal number of pure noise samples (the duration of the signal being kept fixed), the curve should be plotted against m', which is equivalent to moving the curve to the left a distance corresponding to a factor of 2. The resulting curve still lies above the solid curve (for which $m = m'$) for a noise-like signal, but only by about 1 db, which is about the same loss that one finds in Figs. 1 and 2.

It is natural to ask which one should guard against the more, taking too many samples or too few? The penalty for the former increases rather slowly as we have just seen. The latter is a more serious mistake (especially for the noise-like signal of Fig. 3) if one remains limited to a single opportunity to detect the signal. However, if the signal samples missed by one average are included in another and so give another opportunity for detection, the loss is reduced but not eliminated, as shown below in connection with repeated searches.

7. DISTRIBUTION OF A FIXED AMOUNT OF SIGNAL ENERGY

The preceding discussion is closely related to the question of the optimum utilization of a fixed amount of signal energy; assuming that $m = m'$, is it better to have a big pulse of signal occurring in only one sample point, or a lower signal power occurring in a proportionately greater number of samples? Here the product of S/N and m is held constant, and one moves along a 45° line in *all* cases, including Case 3. Reference to the curves shows that with a sinusoidal signal one should take $m = 1$, or at least concentrate the signal energy enough to make the power ratio S/N very much above unity. With a noise-like signal, the same conclusion ($m = 1$) holds under conditions such that $\beta \leqq 50$ per cent, but when better detection is possible a larger value of m may be preferable. For example, take the solid curves ($\alpha = 10^{-6}$) in Fig. 3 and suppose that a one-pulse signal makes $S/N = 18$ db. The corresponding 45° line is tangent to the curve $\beta = 99$ per cent at about $m = 15$. Thus the maximum detection probability is 99 per cent and is attained by taking $S/N = 6.5$ db, and so distributing the signal energy among 15 independent samples. Detection with 99 per cent probability by means of a single sample would require 20 times as much energy.

8. REPEATED SEARCHES

Repeated searches have two important characteristics: (a) There is no carry-over of data from one search to the next. This is not desirable, as we shall see, and would expect on theoretical grounds; but it may be unavoidable. (b) They involve repeated opportunities to detect the signal (or its source).

Repeated searches arise non-trivially when the signal recurs periodically, or can be made to do so by the detecting agency. This assumes that it is sufficient to detect the signal at one of its appearances, so that it is not necessary to regard each appearance as a separate signal. Examples are a signal of long duration but unknown frequency, and the radar return to a search radar from an airplane whose position in space is unknown. Then if one search of frequency or of space does not detect the signal, it may be possible to repeat the search one or more times before the source of the signal disappears.

An equivalent of repeated searches arises when the averaging time is short enough so that two or more averages fall within the duration of the signal. The results below apply to this case and show that a longer average (i.e., carry-over of data from one search to the next) is somewhat preferable. These results do not apply to the case where over-lapping averages give several opportunities to detect the signal, because such averages have some data in common; but it has already been shown that such overlapping is desirable, other things being equal.

Suppose now that the signal power is constant and the time consumed is proportional to the product of m and the number λ of searches. How can one find the signal most quickly? If the probability β of detection in one search is $1/2$, the probability that exactly λ searches are required for detection is $1/2^{\lambda}$, and this also happens to be the probability that the signal remains undetected after λ searches. Thus one would need to make nearly 7 searches before one could conclude with 99 per cent assurance that no signal was present. However, if the signal is in fact present, the *average* number of searches required to detect the signal with 99 per cent assurance is

$$1/2 + 2/2^2 + 3/2^3 + 4/2^4 + 5/2^5 + 6/2^6 + 7/2^6 \doteq 2$$

searches if the superfluous part of the successful search is included, or

$$(1/2)\cdot(1/2) + (3/2)\cdot(1/2^2) + \cdots + (11/2)(1/2^6) + (13/2)(1/2^6) \doteq 3/2$$

searches if the search is terminated the moment the signal is detected.

The latter result is the less favorable for repeated searching, since it is 3 times the result (½) obtained with 99 per cent detection in a single search (which on the average could perhaps be terminated, due to detection of the signal, when it was half completed).

The alternative way of increasing β from 50 per cent to 99 per cent is to increase the value of m by a factor lying between 2.3 and 5 for Figs. 1 and 3, while the fading sinusoid of Figure 2 requires a factor of the order of 1,000. Comparing these with the factors 2 (or 3) and 7 obtained previously gives the following conclusions, which have been confirmed by considering some other values of β:

A. With a fading sinusoidal signal, repeated searching is extremely advantageous, *provided* it is true as assumed that independent samples of the signal are obtainable from search to search but not within a search.

B. Repeated searching may have a small advantage in Cases 1 and 3 (the steady sinusoid and the noise-like signal) provided (1) each search has $\beta \geq 50$ per cent, and (2) the criterion is the average time required to detect the signal when a signal is present, and one is not concerned with the (increased) time required to conclude that no signal is present (which of course is also the time required for the detection of some of the signals).

C. Repeated searching is somewhat disadvantageous in other cases, as one would expect; e.g., in Cases 1 and 3, if a fixed amount of time is available, the greatest probability of detection is achieved by using all of the time for a single search, rather than dividing it among several less sensitive searches.

Mathematical Appendix

9. Pure noise (and noise-like signal = case 3)

If narrow-band Gaussian noise is applied to a square-law rectifier, the value of the output at any instant has a Rayleigh or exponential distribution, as is well known. The average of m such values, taken far enough apart to be virtually independent has a chi-square (χ^2) distribution with $2m$ degrees of freedom. The standard form of the distribution used in tables chooses the units so that the mean of the distribution is $2m$, whereas we want the mean to equal the noise power N. The exact value of k can therefore be found by means of the relation

$$Pr(N\chi^2_{2m}/2m > (k+1)N) = \alpha \tag{9a}$$

or

$$Pr(\chi^2_{2m} > 2m(k+1)) = \alpha$$

where α is the expected ratio of false detections to total number of decisions, when no signal is actually present.

The usual short tables of χ^2 (e.g., that in Fisher and Yates' Statistical Tables or that of Hartley and Pearson in Biometrika **37,** p. 313) are useful but not entirely adequate for the needs of the problem. The most extensive table is that of K. Pearson,[8] which goes to $m = 50$. In terms of the function I tabulated by Pearson,

$$Pr(\chi^2_{2m} > A) = 1 - I(A/2\sqrt{m}, m - 1) \tag{9b}$$

For $m > 50$ it is sufficient to use the expansion[9]

$$\chi^2_{2m} = 2m + 2um^{1/2} + \tfrac{2}{3}(u^2 - 1) + \tfrac{1}{18}(u^3 - 7u)m^{-1/2} - \cdots \tag{9c}$$

Here u is the standard normal[10] deviate. Thus if u_α is defined by

$$\alpha = Pr(u > u_\alpha) = \int_{u_\alpha}^{\infty} e^{-u^2/2}\, du/\sqrt{2\pi}$$

one has

$$k = u_\alpha/\sqrt{m} + (u_\alpha^2 - 1)/3m + \cdots \tag{9d}$$

If a noise-like signal is present, this is mathematically the same as an increase in the average noise power from N to $N + S$, the critical power level remaining at $(k + 1)\,N$. The probability of exceeding the critical level is now the probability β of a true detection. Thus

$$\begin{aligned} \beta &= Pr[(N + S)\chi^2_{2m}/2m > (k + 1)N] \\ &= Pr[\chi^2_{2m} > 2m(k + 1)/(1 + S/N)] \end{aligned} \tag{9e}$$

Writing $\chi^2_{2m}(\beta)$ for the number that is exceeded by the variable χ^2_{2m} with probability β, one gets from (9a) and (9e)

$$1 + S/N = \chi^2_{2m}(\alpha)/\chi^2_{2m}(\beta) \tag{9f}$$

For large m (and only in that case), (9c) then gives

$$S/N \doteq (u_\alpha - u_\beta)\left[\frac{1}{\sqrt{m}} + \frac{u_\alpha - 2u_\beta}{3m}\right] \tag{9g}$$

[8] K. Pearson, Tables of the Incomplete Gamma-Function, Biometrika Office, London, 1934.

[9] The author first discovered this formula in T. Lewis, Biometrika **40,** p. 424, 1953; but S. O. Rice has pointed out that G. A. Campbell published it as early as 1923 in the B.S.T.J., **2,** page 95, in connection with Poisson distribution. Other references and inversion formulas are given by John Riordan. Ann. Math. Stat., **20,** p. 417, 1949.

[10] "Normal" is a synonym for "Gaussian." However a new application of the distribution is involved here, and so there is no disadvantage in making the conventional change in terminology.

where the standard normal deviate u_β is defined by $\beta = Pr(u > u_\beta)$. It is negative when $\beta > 50$ per cent.

Wherever it occurs in equations, S/N naturally means the ratio of S to N, and not the value in db, which is $10 \log_{10} S/N$.

10. NOISE PLUS STEADY SINUSOID (CASE 1)

When a pure sinusoidal signal of constant amplitude is added to the noise, the distribution of the rectifier output has what is called a non-central chi-square distribution, which has been little tabulated. Fortunately the normal distribution gives a fair approximation in this case (with the probable exception of small values of β, which are of little interest anyway). More accuracy could be obtained at the cost of additional labor.[11]

Before rectification, the signal plus noise can be represented, as remarked in Section 3, by the expression

$$\sqrt{2S} \cos(\Omega t + \varphi) + n_1(t) \cos(\Omega t + \varphi) + n_2(t) \sin(\Omega t + \varphi) \quad (10a)$$

where the two Gaussian noise variables $n_1(t)$ and $n_2(t)$ are independent and have zero means and a common variance N. After rectification and smoothing, half the square of the envelope is obtained, namely

$$[(\sqrt{2S} + n_1(t))^2 + n_2(t)^2]/2$$

Its mean value is $N + S$, and its variance is the sum of the variances of the terms in the expanded form, namely $0 + 2NS + N^2/2 + N^2/2$ or $2NS + N^2$. The average of m such independent variables has the same mean $N + S$ but the variance $N(N + 2S)/m$.

The critical value $(k + 1)N$ of the rectifier output is reduced to standard units by subtracting the mean output $N + S$ and dividing by the square root of the variance, giving

$$u_\beta = (k - S/N) \sqrt{\frac{m}{1 + 2S/N}}$$

After u_β is found from

$$\beta = Pr(u > u_\beta) = \int_{u_\beta}^{\infty} e^{-u^2/2}\, du/\sqrt{2\pi}$$

by using a table of the normal distribution, one can solve the preceding

[11] P. B. Patnaik, Biometrika **36**, p. 202, 1949.

equation for S/N, giving

$$\frac{S}{N} \doteq k + \frac{u_\beta^2}{m} - \frac{u_\beta}{\sqrt{m}}\left[1 + 2k + \frac{u_\beta^2}{m}\right]^{1/2} \tag{10b}$$

This is then the signal-to-noise ratio giving the probability β of detection.

By using (9d) and taking $1 + u_\alpha/\sqrt{m}$ as an approximate value of the radical, one obtains from (10b)

$$\frac{S}{N} \doteq \frac{u_\alpha - u_\beta}{\sqrt{m}} + \frac{u_\beta(u_\beta - u_\alpha) + (u_\alpha^2 - 1)/3}{m} \tag{10c}$$

This result differs most (a little over 1 db) from the curves of Figure 1 when $m = 1$ and $\beta = 0.01$, but (10b) itself is not very accurate in that case.

11. NOISE PLUS FADING SINUSOID (CASE 2)

In this case the signal plus noise has the form

$$[s + n(t)] \cos \Omega t + [s' + n'(t)] \sin \Omega t \tag{11a}$$

where $n'(t)$ is *not* a derivative, and the components s and s' of the signal amplitude are essentially constant during the period that is averaged over, although like n and n' they are Gaussian variables of mean zero. Detection is accomplished by means of the expression

$$R = \frac{1}{2m}\left[\sum_1^m (s + n_i)^2 + \sum_1^m (s' + n_i')^2\right] \tag{11b}$$

where the n_i and n_i' are (independent) values of $n(t)$ and $n'(t)$ at successive sampling points.

Now (11b) can be written

$$\begin{aligned} 2R = &\left(s + \frac{1}{m}\sum_1^m n_i\right)^2 + \left(s' + \frac{1}{m}\sum_1^m n_i'\right)^2 \\ &+ \frac{1}{m}\sum n_i^2 - \left(\frac{1}{m}\sum n_i\right)^2 + \frac{1}{m}\sum n_i'^2 - \left(\frac{1}{m}\sum n_i'\right)^2 \end{aligned} \tag{11c}$$

as in the analysis of variance in statistics. The sum of two squares in the first line of (11c) is then distributed as $(S + N/m)\chi_2^2$; the second line is distributed as $\chi_{2m-2}^2 \cdot N/m$; and these portions are independent of one another, because they correspond to the mean and the variance respec-

tively of the sample of n_i (or n_i'), and the sample mean and variance are known to be independent in the Gaussian case. Here the signal power S is the variance of each of s and s', and N is the variance of each of the n_i and n_i'.

For detection R must exceed $N(1 + k)$. Hence if S/N is denoted by r, one has

$$\beta = Pr[(1 + mr)\chi_2^2 + \chi_{2m-2}^2 > 2m(1 + k)] \tag{11d}$$

Since $(1 + mr)\chi_2^2$ has the c.d.f. $1 - e^{-t/2(1+mr)}$ and χ_{2m-2}^2 has the density function

$$e^{-t/2}t^{m-2}/2^{m-1}\Gamma(m - 1),$$

β is obtained by convolution as

$$\begin{aligned}\beta &= 1 - \int_0^z [1 - e^{-(z-t)/2(1+mr)}]e^{-t/2}t^{m-2}\,dt/2^{m-1}\Gamma(m - 1) \\ &= Pr[\chi_{2m-2}^2 > z] + e^{-z/2(1+mr)}\left(\frac{mr}{1 + mr}\right)^{1-m} Pr\left[\chi_{2m-2}^2 < \frac{mrz}{1 + mr}\right]\end{aligned} \tag{11e}$$

where $z = 2m(1 + k)$.

An excellent approximation is obtained by replacing the variable χ_{2m-2}^2 by its mean value $2m - 2$; this is reasonable, since the ratio of the variances of χ_{2m-2}^2 and $(1 + mr)\chi_2^2$ is $(m - 1)/(1 + m\ S/N)^2$, which is very small throughout the interesting area ($\beta \geqq 50$ per cent). (11d) then gives

$$\beta \doteq \exp\,[-(1 + km)/(1 + m\ S/N)] \tag{11f}$$

This result will be generalized in (13k) below. Using (9d) then gives

$$\frac{S}{N}\,\ell n\,(1/\beta) \doteq \frac{u_\alpha}{\sqrt{m}} + \frac{(u_\alpha{}^2 + 2)/3 - \ell n\,(1/\beta)}{m} \tag{11g}$$

to a little lower order of accuracy when m is small, and still with $\beta \geqq 50$ per cent.

Before the foregoing results were obtained, the curves of Fig. 2 had already been calculated[12] on IBM equipment by the rather tedious process of integrating the results for Case 1 with respect to the random signal

[12] By Mrs. L. R. Lee, at the request of the author.

strength. This gives

$$\beta \doteq \int_0^\infty \Phi[(xS/N - k)\sqrt{m/(1 + 2xS/N)}]p(x)\,dx \qquad (11h)$$

for Case 2, where

$$p(x) = e^{-x}$$

and

$$\Phi(u) = \int_{-\infty}^{u} e^{-t^2/2}\,dt/\sqrt{2\pi}$$

If the fading is due to interference between two sinusoidal signals of powers S_1 and $S_2(S_1 + S_2 = S)$ and very slightly different frequencies, it can be shown that the appropriate form for $p(x)$ is $1/\pi\sqrt{A^2 - (1 - x)^2}$ for $1 - A < x < 1 + A$, and zero elsewhere, where $A = 2\sqrt{S_1S_2}/(S_1 + S_2)$. Here x can approach zero only in the special case $S_1 = S_2$. Since one or both of S_1 and S_2, and hence A, is likely to be a random variable, one apparently cannot say a priori that any particular distribution $p(x)$ is the appropriate one in this case.

12. COMPARISON OF 12 CASES

The three principal cases defined in Section 2 and analyzed in the last three sections may be expanded to 12 by considering dc signals as well as ac, and considering also a second method of detection. The original 3 cases will be labeled "ac envelope." (More precisely, half of its square is the quantity averaged.) If the envelope is not isolated but instantaneous squared values are sampled, the situation is "ac instantaneous." A precise analysis of these latter cases would apparently be difficult. As indicated, quadratic rectification is assumed for all 6 of the ac cases. On the other hand, for the 6 dc cases, distinguished as rectified and unrectified, it is the instantaneous sampling that is assumed throughout.

The reduction of the rectified dc cases to the standard ac envelope cases has been indicated at the end of Section 3, and will be demonstrated at the beginning of the following section. The unrectified dc cases, which represent the better way to detect a steady or fading dc signal, involve nothing but the simple Gaussian distributions specified in Table I. If the number of samples averaged is denoted by $2m$, one easily finds the theoretically exact relations

$$S/N = (u_\alpha - u_\beta)^2/2m \qquad (12a)$$

$$S/N = (u_{\alpha/2}^2/u_{\beta/2}^2 - 1)/2m \qquad (12b)$$

$$S/N = u_{\alpha/2}^2/u_{\beta/2}^2 - 1 \qquad (12c)$$

TABLE I — VARIABLES TO BE AVERAGED

	DC Unrectified	DC Rectified
Steady	$\sqrt{S} + n$	$S + 2n\sqrt{S} + n^2$
Fading Noise-like	$s + n$	$s^2 + 2ns + n^2$
	AC Envelope	
Steady	$S + n\sqrt{2S} + (n^2 + n'^2)/2$	
Fading Noise-like	$[(s + n)^2 + (s' + n')^2]/2$	
	AC Instantaneous	
Steady	$[(\sqrt{2S} + n)\cos\varphi + n'\sin\varphi]^2$	
Fading	$[(s + n)\cos\varphi + (s' + n')\sin\varphi]^2$	
Noise-like	Ditto or $s^2 + 2ns + n^2$	

for the steady, fading, and noise-like unrectified dc (or low-pass) signals. As usual, u_p is defined by

$$\int_{u_p}^{\infty} e^{-t^2/2}\, dt/\sqrt{2\pi} = p$$

In (12a) it is assumed that the steady dc signal of magnitude $\sqrt{S}$ is of known sign. Otherwise it would be necessary, as with fading and noise-like signals, to be prepared to detect both positive and negative dc signals, and quantities like $u_{\alpha/2}$ enter in place of u_α via the relation

$$\int_{u_{\alpha/2}}^{\infty} + \int_{-\infty}^{-u_{\alpha/2}} = \alpha.$$

As β approaches unity, $u_{\beta/2}$ approaches zero and S/N rapidly approaches infinity for the fading and the noise-like signal. In the latter case (12c), averaging has no effect (m does not appear), because the unrectified signal averages to zero as fast as the noise does. In fact, with instantaneous sampling of a noise-like signal, there is no essential difference between the lowpass ("dc") and narrow-band (ac) cases.

If n, n', s, s' are independent Gaussian variables with zero means and variances N, N, S, S respectively, and φ is uniformly distributed between 0 and 2π, the variables that are averaged in the detection process have the forms given in Table I in the various cases, when a signal is present (if not, put s, s', and S equal to zero).

Table II gives the variances of the variables of Table I for the cases

TABLE II — VARIANCES

	DC Unrectified	DC Rectified
Steady	N	$2N^2 + 4NS$
Noise-like	$N + S$	$2N^2 + 4NS + 2S^2$
	AC Envelope	**AC Instantaneous**
Steady	$N^2 + 2NS$	$2N^2 + 4NS + \frac{1}{2}S^2$
Noise-like	$N^2 + 2NS + S^2$	$2N^2 + 4NS + 2S^2$

of a steady or a noise-like signal. The means are $\sqrt{S}$ and zero, respectively, for the dc unrectified variables, while all the others have the power $N + S$ as their mean.

Table III gives the data for the fading signal. The reducible variance and the mean are calculated under the condition that s and s' are fixed; the irreducible variance is then the variance of the resulting mean when s and s' do vary. The distinction between the two variances is that by averaging, the first is reduced in the usual way (divided by the number of samples), while the second retains its full value. The variance already given for the steady signal can be derived from the reducible variances above by replacing s^2 and s'^2 by S. The variances already given for the noise-like signal are equal to the irreducible variance plus the expected value of the reducible variance.

The steady ac instantaneous case will serve to illustrate the calculation of the variance. If the rectifier output is written in the form

$$S + n\sqrt{2S} + (n^2 + n'^2)/2 + (n'\sqrt{2S} + nn') \sin 2\varphi + [S + n\sqrt{2S} + (n^2 - n'^2)/2] \cos 2\varphi$$

it may be verified that the various terms are uncorrelated with one another, so that their individual variances may be added. Also $E \sin 2\varphi = E \cos 2\varphi = 0$, $E \sin^2 2\varphi = E \cos^2 2\varphi = \frac{1}{2}$, $En^4 = En'^4 = 3N^2$,

TABLE III — FADING SIGNALS

	Reducible Variance	Mean	Irreducible Variance
DC unrectified	N	s	S
DC rectified	$2N^2 + 4Ns^2$	$N + s^2$	$2S^2$
AC envelope	$N^2 + N(s^2 + s'^2)$	$N + (s^2 + s'^2)/2$	S^2
AC instantaneous	$2N^2 + 2N(s^2 + s'^2) + (s^2 + s'^2)^2/8$	$N + (s^2 + s'^2)/2$	S^2

var n^2 = var $n'^2 = 2N^2$. These last relations are not obvious, but are a special case ($t = u$) of the relation cited below just prior to (15e). The variance is now obtained as

$$0 + 2NS + 4N^2/4 + \tfrac{1}{2}E(2Sn'^2 + n^2n'^2) + \tfrac{1}{2}E[S^2 + 2Sn^2 + (n^4 - 2n^2n'^2 + n'^4)/4] = 2N^2 + 4NS + S^2/2$$

The variances tabulated indicate the relative merit of the various cases. If m is significantly greater than unity, the anomalous case of an unrectified noise-like signal represented by (12c) gives the poorest detection; next come the fading signals with their irreducible variances. The steady dc unrectified signal with S/N proportional to m^{-1} (by 12a) is the easiest to detect. The steady signals are in all cases more easily detected than the noise-like at the higher signal strengths. However, instantaneous sampling of a steady ac signal loses some of this advantage; a term $S^2/2$ appears, in addition to the expected doubling of the variance of a single sample.

13. DETECTION OF AN ARBITRARY GAUSSIAN SIGNAL

General formulas will now be derived which include Cases 1, 2, and 3 and also give approximations to intermediate cases. It will be more convenient to deal with the Gaussian sample values $s_i + n_i$ of the signal plus noise in the dc or lowpass case, and $2m$ of them will be taken so that the results obtained will have the same form as those already given for the ac or narrow-band case. Detection is then accomplished by means of the variable

$$R = \sum_1^{2m} (s_i + n_i)^2/2m = \sum_1^{2m} (s_i^2 + 2s_in_i + n_i^2)/2m \qquad (13a)$$

It is assumed that the n_i are independent of one another and of the s_i, while the s_i may be constant or mutually correlated. Also $En_i = 0$ and $En_i^2 = N$, a constant.

An orthogonal transformation (rotation in $2m$-dimensional space) can be used to show that (13a) has the same distribution as

$$\sum_1^{2m} (h + n_i')^2/2m \qquad (13b)$$

or

$$\left[\sum_1^{m} (h\sqrt{2} + n_i'')^2 + \sum_{m+1}^{2m} n_i''^2\right] \Big/ 2m \qquad (13c)$$

where $h^2 = \sum_1^{2m} s_i^2/2m$, and the n_i' or n_i'' are new noise variables with exactly the same properties as the original n_i. Evidently h^2 is so defined as to leave unchanged the terms independent of the noise variables, while the orthogonal transformation by definition makes $\sum n_i^2 = \sum n_i'^2 = \sum n_i''^2$. Thus it is only necessary to choose the transformation so that the original linear terms $\sum_1^{2m} s_i n_i/m$ go into $h \sum_1^{2m} n_i'/m$ or $\sqrt{2}h \sum_1^m n_i''/m$. This is always possible since all three expressions have the same norm (square root of sum of squares of coefficients), namely $h\sqrt{2/m}$.

The forms (13b) and (13c) are appropriate to the instantaneous sampling of the square of a dc signal plus noise, and the sampling of the square of the envelope of an ac signal plus narrow-band noise, respectively, the latter being the standard case. Thus the equivalence of the two cases is established. It is only necessary to observe that the number of samples taken is $2m$ in (13b) but only m in (13c). The noise powers are En_i^2 and

$$E(n_i \cos\varphi + n_{m+i} \sin\varphi)^2 = E(n_i^2 + n_{m+i}^2)/2 = En_i^2.$$

The signal powers are h^2 and $E \cdot 2h^2 \cos^2\varphi = h^2$. Another consequence of (13b) or (13c) is that when the distribution of h has a known form, the distribution of R could be obtained from the results for Case 1 by integrating over h, as stated at the end of Section 11.

The mean and the variance of the general form (13a) will now be calculated. It is assumed that $Es_i = 0$ and $Es_i^2 = S$ for the first $2m'$ values of $i (m' \leqq m)$, while $s_i \equiv 0$ for the other $2(m - m')$ values of i. Then one has $ER = N + m'S/m$. If i and j are any two distinct integers, the five variables s_i^2, $s_i n_i$, $s_j n_j$, n_i^2, n_j^2 are all uncorrelated (though not all independent), and so one has

$$4m^2 \text{ var } R = \text{var } (\sum s_i^2) + 4 \sum \text{var } (s_i n_i) + 2m \text{ var } n_i^2$$

Since var $(s_i n_i) = NS$ or 0 and var $n_i^2 = 2N^2$, this gives

$$\text{var } R = \text{var } h^2 + 2m'NS/m^2 + N^2/m \tag{13d}$$

with $h^2 = \sum_1^{2m'} s_i^2/2m$. In Cases 1, 2, and 3, var $h^2 = 0$, $(m'S/m)^2$, and $m'S^2/m^2$ respectively. (In Case 2, m' of the s_i have one identical value and m' have another identical value, independent of the first.) In general

$$\text{var } h^2 = \sum_{-2m'+1}^{2m'-1} (2m' - |i|)\psi^2(i)/2m^2 \tag{13e}$$

where $\psi(i) = Es_j s_{i+j}$ for all j for which $s_j s_{i+j} \neq 0$. This is the discrete analogue of the integral appearing in (15k) below. It is convenient to define a number m'' (which need not be an integer) such that var $h^2 =$

$(m'S/m)^2/m''$. Then putting $S' = m'S/m$ gives

$$ER = N + S', \qquad \text{var } R = S'^2/m'' + (2NS' + N^2)/m. \quad (13f)$$

The mean and variance of R are identical with those of $[S'^2/m'' + (2NS' + N^2)/m]/2(N + S')$ times a chi-square variable with

$$2\bar{m} = 2(N + S')^2/[S'^2/m'' + (2NS' + N^2)/m]$$

degrees of freedom. So one has approximately

$$\beta = Pr(R > (k + 1)N)$$
$$\doteq Pr\{\chi^2_{2\bar{m}} > 2(k + 1)N(N + S')/[S'^2/m'' + (2NS' + N^2)/m]\} \quad (13g)$$
$$\doteq Pr\{u > (kN - S')/\sqrt{S'^2/m'' + (2NS' + N^2)/m}\}$$

The last line can be used to verify that for fixed m, the optimum value of m' is m, and for fixed m', the optimum value of m is m'. Defining u_β by $Pr(u > u_\beta) = \beta$ and solving for $S/N = mS'/m'N$ gives

$$S/N = \frac{mk + u_\beta^2 - u_\beta\sqrt{(2k + 1)m + k^2m^2/m'' - u_\beta^2(m/m'' - 1)}}{m'(1 - u_\beta^2/m'')} \quad (13h)$$

The approximation breaks down if $u_\beta^2 \geqq m''$.

It can he shown that by replacing the symbols m', m'', and S in the above formulas by $m'/2$, $m''/2$, and $2S$, one obtains the results for the narrow-band case in which the signal is a noise-modulated carrier $s(t)$ cos Ωt, while the noise as usual has the form $n(t)$ cos $\Omega t + n'(t)$ sin Ωt.

In Cases 1, 2, and 3, $m'' = \infty$, 1, and m' respectively; also, m' has been assumed equal to m. In the dc or lowpass case, m'' could be as small as $\frac{1}{2}$. One has $m \geqq m'$ always, and $m' \geqq m''$ when the signal is Gaussian (i.e., contains no steady sinusoidal or steady dc component).

When m'' is very small, (13g) and (13h) are of low accuracy, and it is much better to use (11e) or (11f), or the following generalization of them. The derivation proceeds as if m'/m'' were an integer, although this is probably not a necessary condition for the usefulness of the results. The averaged rectifier output is represented by

$$2mR = \sum_{i=1}^{2m''} \sum_{j=1}^{m'/m''} (s_i + n_{ij})^2 + \sum_{1}^{2m-2m'} n_i^2 \quad (13i)$$

where the n_{ij} are $2m'$ independent Gaussian noise variables accompanying the signal variables s_i (which have only $2m''$ independent values), and the n_i are $2m - 2m'$ additional noise variables that are not accom-

panied by a signal. Of course, the signal variables would not in reality fall into independent sets of identical values, so that an approximation enters here even if m'/m'' is an integer.

(13i) may be written as

$$2mR = (m'/m'') \sum_{i=1}^{2m''} (s_i + \bar{n}_{i+})^2 + \sum_{i=1}^{2m''} \sum_{j=1}^{m'/m''} (n_{ij} - \bar{n}_{i+})^2 + \sum_{1}^{2m-2m'} n_i^2 = (N + m'S/m'')\chi^2_{2m''} + N\chi^2_{2m-2m''} \tag{13j}$$

where

$$\bar{n}_{i+} = (m''/m') \sum_{j=1}^{m'/m''} n_{ij}$$

$$S = \text{var } s_i$$

$$N = \text{var } n_{ij} = \text{var } n_i$$

and the two χ^2 variables are independent. If the ratio $m''(m - m'')/(m'' + m'S/N)^2$ of the variances of the latter (including their multipliers) is small, one may replace $\chi^2_{2m-2m''}$ by its mean value $2m - 2m''$ and obtain

$$\begin{aligned} \beta &= Pr[R > (1 + k)N] \\ &\doteq Pr[\chi^2_{2m''} > 2(m'' + mk)/(1 + m'S/m''N)] \end{aligned} \tag{13k}$$

14. COUNTING SAMPLES ABOVE A THRESHOLD

It is sometimes suggested that instead of averaging the m samples, the number of such samples exceeding some threshold might be counted and used as the detection criterion. This is equivalent to replacing the average of the m samples by one of their order statistics, such as the median.

It is of interest to ask which order statistic is best to use, and how it compares in efficiency with the average. M. Schwartz (in the dissertation cited previously) made numerical calculations for the case of a steady sinusoid and $m \leqq 49$, and concluded that the method of coincidences, as he called it, required S/N to be about 1.4 db above that which sufficed for equal performance using averages. For the larger values of m there appeared to be a small advantage in requiring less than half the samples to exceed the (suitably chosen) threshold. These results are confirmed by the following asymptotic analysis.

In the absence of a signal, a single sample of half the square of the

envelope has an exponential distribution $e^{-x/N}\,dx/N$. The same (with increased power N) is true in the presence of a noise-like signal, and approximately so for a sinusoidal signal of low intensity. Since m is assumed to be fairly large, low-intensity signals are the interesting ones. In both of these cases detection can he based on an estimate of N, denoted by $\hat{N}$.

In the present method $\hat{N}$ is determined from $\hat{p} = e^{-K/\hat{N}}$, where $\hat{p}$ is the proportion of samples observed to exceed the threshold K. The standard deviation σ_p of $\hat{p}$ is well-known to be $\sqrt{p(1-p)/m}$, where $p = e^{-K/N}$ = the "true" or expected value of $\hat{p}$. For large m the sampling fluctuations are small and the standard deviation σ'_N of $\hat{N}$ is approximately equal to σ_p times dN/dp or $\sigma_p \div e^{-K/N} \cdot K/N^2$ or

$$\sigma'_N = N(1-p)^{1/2}/(mp)^{1/2}\,\ell n\,(1/p) \tag{14a}$$

for counting samples as compared with

$$\sigma_N = N/\sqrt{m}$$

when the average is used. So the efficiency of the counting procedure (expressed in terms of the (inverse) number of samples required for equivalent reliability) is

$$\sigma_N^2/\sigma_N'^2 = (\ell n\ 1/p)^2 p/(1-p) \tag{14b}$$

This expression has its maximum value of 64.7 % when $p = e^{-2(1-p)} \doteq 0.203$. The median ($p = \frac{1}{2}$) has an efficiency of only 48.0 %. For large m, the required S/N varies as $m^{-1/2}$ and so the minimum loss due to counting is $10 \log_{10} 0.647^{-1/2} \doteq 1.0$ db.

In the case of the steady unrectified dc signal, the detection problem is equivalent to estimating the mean μ of a Gaussian distribution

$$\varphi[(x-\mu)/\sigma]\,dx = e^{-(x-\mu)^2/2\sigma^2}\,dx/\sigma\sqrt{2\pi}.$$

The estimate $\hat{\mu}$ of μ is determined from

$$\hat{p} = 1 - \Phi\left(\frac{K-\hat{\mu}}{\sigma}\right) = \int_K^\infty e^{-(x-\hat{\mu})^2/2\sigma^2}\,dx/\sigma\sqrt{2\pi}$$

and the same relation holds between the true values p and μ. Then $dp/d\mu = \varphi[(K-\mu)/\sigma]/\sigma$ and

$$\sigma'_\mu = \sigma\sqrt{\Phi(\xi)[1-\Phi(\xi)]}/\varphi(\xi)\sqrt{2m} \tag{14c}$$

where $\xi = (K-\mu)/\sigma$ and $2m$ is the number of samples. The variance σ_μ of the mean is $\sigma/\sqrt{2m}$, so the efficiency of the counting procedure is

$$\sigma^2_\mu/\sigma'^2_\mu = \varphi^2(\xi)/\Phi(\xi)[1-\Phi(\xi)] \tag{14d}$$

The following are some values of this expression:

ξ	0	0.5	1	2
Efficiency	63.7%	58.0%	43.9%	13.1%

Thus the highest efficiency of 63.7% is attained when $\xi = 0$ or the threshold K equals the mean μ. This corresponds to using the median as the statistic on which detection is based. This maximum efficiency is very nearly the same as that obtained in the preceding case. However, in the present case S/N varies as m^{-1} by (12a), so that the equivalent loss in signal strength is not 1.0 but $10 \log_{10} 1/0.637 \doteq 2.0$ db.

15. AVERAGING BY CONTINUOUS INTEGRATION

In practice, in place of the discrete sums of squares such as (11b) and (13a) one may have integrals such as

$$R_U = \int_0^T Z(t)\, dt/T \qquad \text{and} \tag{15a}$$

$$R_E = \int_{-\infty}^0 Z(t) e^{t/T}\, dt/T \tag{15b}$$

where $Z(t)$ is the variable to be averaged (usually a rectifier output), and the subscripts U and E refer to uniform and exponential weighting respectively. The purpose of this section is to calculate the mean μ and variance σ^2 of each of these expressions for various cases. It will then be shown that for a flat spectrum and $T \to \infty$, μ and σ^2 are the same as those already given for the discrete averages of samples. The results are closely related to those given by Rice.[13]

It will suffice to consider steady signals. If the signal is absent, put $S = 0$. If the signal is fading, S is a random variable. If the signal is noise-like (with the same spectrum as the noise), put $S = 0$ and replace N by $N + S$. The four combinations of ac and dc signals with two methods of detection (see Section 12) will be considered separately.

The exceptional case of an unrectified dc signal may be disposed of first. Then $Z(t) = \sqrt{S} + n(t)$ and R_U and R_E both have Gaussian distributions with mean $\sqrt{S}$. The variance of R_U is

$$E\left(\int_0^T n(t)\, dt/T\right)^2$$

Writing the square as a double integral and taking the expectation under

[13] S. O. Rice, B.S.T.J., **24**, pp. 46–156, 1945, Eqns. 3.9-8 and 3.9-28. Also J. Acous. Soc. of Amer., **14**, pp. 216–227, 1943.

the integral sign gives

$$\begin{aligned} \text{var } R_U &= E\int_0^T\int_0^T n(t)n(u)\,dt\,du/T^2 \\ &= \int_0^T\int_0^T \psi(t-u)\,dt\,du/T^2 = 2\int_0^T (T-v)\psi(v)\,dv/T^2 \end{aligned} \tag{15c}$$

where $\psi(v) = En(t)n(t+v)$, and $\psi(0) = N$. Similarly

$$\begin{aligned} \text{var } R_E &= \int_0^\infty\int_0^\infty e^{-(t+u)/T}\psi(t-u)\,dt\,du/T^2 \\ &= \int_0^\infty e^{-v/T}\psi(v)\,dv/T \end{aligned} \tag{15d}$$

In the remaining three cases, R_U and R_E always have the power $N + S$ as their mean. Their distributions are not known exactly but they might be assumed to be distributed approximately like $(\sigma^2/2\mu)\chi^2_{2\mu^2/\sigma^2}$ where μ is the mean $N + S$ and σ^2 is the appropriate variance given below. A probably more convenient procedure is to use the $\mu = N$ and $\sigma^2 = \sigma_0^2$ (see 15k and ℓ) for the noise alone to determine an equivalent value N^2/σ_0^2 for m, and then use Figs. 1–3.

Consider next the rectified dc signal, so that

$$Z(t) = S + 2\sqrt{S}n(t) + n^2(t)$$

The variance of R_U is $E[R_U - S - N]^2$ or

$$E\left(\int_0^T [n^2(t) - N]\,dt/T + 2\sqrt{S}\int_0^T n(t)\,dt/T\right)^2.$$

The cross-product has zero expectation. Expressing the squares as double integrals gives

$$E\int_0^T\int_0^T ([n^2(t) - N][n^2(u) - N] + 4Sn(t)n(u))\,dt\,du/T^2.$$

We now take the expectation under the integral sign. For

$$En^2(t)n^2(u) = \psi^2(0) + 2\psi^2(t-u)$$

see M. G. Kendall, The Advanced Theory of Statistics, Volume I, Section 3.28, equation

$$\mu_{22} = (1 + 2\rho^2)\sigma_1^2\sigma_2^2$$

This gives

$$\begin{aligned} \text{var } R_U &= \int_0^T \int_0^T [2\psi^2(t-u) + 4S\psi(t-u)]\, dt\, du/T^2 \\ &= 4\int_0^T (T-v)[\psi^2(v) + 2S\psi(v)]\, dv/T^2 \end{aligned} \tag{15e}$$

A similar treatment of R_E gives

$$\text{var } R_E = 2\int_0^\infty e^{-v/T}[\psi^2(v) + 2S\psi(v)]\, dv/T. \tag{15f}$$

The third case is that of a sinusoidal signal with instantaneous sampling, for which

$$Z(t) = 2S\cos^2(\Omega t + \varphi) + 2\sqrt{2S}n(t)\cos(\Omega t + \varphi) + n^2(t)$$

where φ is uniformly distributed in $(0, 2\pi)$. The variance of R_U is now given by

$$\begin{aligned} E\Bigg(\int_0^T [n^2(t) - N]\, dt/T + 2\int_0^T \sqrt{2S}\cos(\Omega t + \varphi)n(t)\, dt/T \\ + \int_0^T S[2\cos^2(\Omega t + \varphi) - 1]\, dt/T\Bigg)^2 \\ = E\int_0^T \int_0^T [n^2(t) - N][n^2(u) - N]\, dt\, du/T^2 \\ = E\cdot 8S\int_0^T \int_0^T \cos(\Omega t + \varphi)\cos(\Omega u + \varphi)n(t)n(u)\, dt\, du/T^2 \\ + E\cdot S^2\int_0^T \int_0^T \cos(2\Omega t + 2\varphi)\cos(2\Omega u + 2\varphi)\, dt\, du/T^2 \end{aligned}$$

To find $E\cos(\Omega t + \varphi)\cos(\Omega u + \varphi)$, expand the cosines, note that $E\cos^2\varphi = E\sin^2\varphi = \frac{1}{2}$ and $E\cos\varphi\sin\varphi = 0$, and combine the resulting terms. This gives

$$\begin{aligned} \text{var } R_U = \int_0^T \int_0^T (2\psi^2(t-u) + 4S\psi(t-u)\cos\Omega(t-u) \\ + \tfrac{1}{2}S^2\cos 2\Omega(t-u))\, dt\, du/T^2 \end{aligned} \tag{15g}$$

$$= 4\int_0^T (T-v)[\psi^2(v) + 2S\psi(v)\cos\Omega v]\, dv/T^2 + (S^2/2T^2\Omega^2)\sin^2\Omega T.$$

A similar treatment of R_E gives

$$\text{var } R_E = 2\int_0^\infty e^{-v/T}[\psi^2(v) + 2S\psi(v)\cos \Omega v]\,dv/T + S^2/2(1 + 4T^2\Omega^2) \tag{15h}$$

of which the last term is the evaluation of

$$\frac{S^2}{2}\int_0^\infty\int_0^\infty e^{-(t+u)/T}\cos 2\Omega(t-u)\,dt\,du/T^2$$

It remains to consider the narrow-band case, with filtering after rectification to give half the square of the envelope. As in Sections 3 and 10, the final output has the form

$$Z(t) = S + \sqrt{2S}n_1(t) + [n_1^2(t) + n_2^2(t)]/2$$

The variance of R_U is then

$$E\left(\frac{1}{2}\int_0^T [n_1^2(t) + n_2^2(t) - 2N]\,dt/T + \sqrt{2S}\int_0^T n_1(t)\,dt/T\right)^2$$

The usual method of evaluation gives

$$\text{var } R_U = 2\int_0^T (T - v)[\psi_0^2(v) + \psi_1^2(v) + 2S\psi_0(v)]\,dv/T^2 \tag{15i}$$

$$\text{var } R_E = \int_0^\infty e^{-v/T}[\psi_0^2(v) + \psi_1^2(v) + 2S\psi_0(v)]\,dv/T \tag{15j}$$

where

$$\psi_0(v) = En_1(t)n_1(t+v) = En_2(t)n_2(t+v)$$

and

$$\psi_1(v) = En_1(t)n_2(t+v) = -En_1(t+v)n_2(t).$$

Then $T \to \infty$, the ratio of the variances of R_E and R_U approaches one-half in all cases, so it will suffice to consider the latter. The spectrum of the noise will be assumed to be flat, of width ω radians per second. Then in (15e) and (15e) one has $\psi(v) = N \sin \omega v/\omega v$, and so the variance of R_U is asymptotically equal to

$$\frac{2}{T}\int_0^\infty \frac{N \sin \omega v}{\omega v}\,dv$$

and

$$\frac{4}{T}\int_0^\infty \left[\frac{N^2 \sin^2 \omega v}{(\omega v)^2} + \frac{2NS \sin \omega v}{\omega v}\right]dv$$

respectively.

Putting $\psi(v) = (2N/\omega v) \sin(\omega v/2) \cos \Omega v$ in (15g) gives

$$\frac{4}{T}\int_0^\infty \left[\frac{N^2 \sin^2 \omega v/2}{(\omega v/2)^2} + \frac{2NS \sin \omega v/2}{\omega v/2}\right] \cos^2 \Omega v \, dv$$

Putting

$$\psi_1(v) = 0, \qquad \psi_0(v) = (2N/\omega v) \sin(\omega v/2)$$

in (15i) gives

$$\frac{2}{T}\int_0^\infty \left[\frac{N^2 \sin^2 \omega v/2}{(\omega v/2)^2} + \frac{2NS \sin \omega v/2}{\omega v/2}\right] dv$$

The first integral has the value $\pi N/\omega T$ and the last three all have the value $(2\pi/\omega T)(N^2 + 2NS)$. Comparing these with the variances of single samples given for steady signals in Table II (Section 12), and writing $\omega/2\pi = W$ cps we see that R_U is asymptotically equivalent to $2WT$ independent samples in the two dc cases, and WT in the ac envelope case. The number is something above $2WT$ for instantaneous samples of ac. These results agree with those arrived at in Section 3 and 13. With the exception of the unrectified dc signal, the differences are seen to lie in the efficacy of an isolated sample, rather than in the long-term rate of transport of information.

When the signal is absent, (15e) to (15h) reduce to the results of Rice cited above:

$$\text{var } R_U = 4\int_0^T (T - v)\psi^2(v) \, dv/T^2 \tag{15k}$$

and

$$\text{var } R_E = 2\int_0^\infty e^{-v/T}\psi^2(v) \, dv/T \tag{15ℓ}$$

The same is nearly true for (15i) and (15j) also, since by (3c), $\psi_0^2(v) + \psi_1^2(v)$ is the square of the envelope of $\psi^2(v)$ in the narrow-band (ac) case.

16. OPTIMUM PROCEDURES AND THE BASIC ASSUMPTIONS

The rigorous determination of the optimum detection procedure is a deep problem with more or less complicated answers depending on the spectra and probably involving both discrete sampling and continuous integration, or even differentiation.[14-16] However, the solution of the prob-

[14] U. Grenander, Stochastic Processes and Statistical Inference, Arkiv för Matematik, **1**, p. 195, 1950, Sections 4.11 and 5.4.

[15] E. Reich and P. Swerling, The Detection of a Sine Wave in Gaussian Noise. J. Appl. Phys., **24**, p. 289, 1953.

[16] D. Slepian, Estimation of Signal Parameters in the Presence of Noise. Trans. I.R.E., Professional Group on Information Theory, p. 68, March, 1954.

lem is known as soon as one restricts oneself to independent samples of the signal-plus-noise. Then square-law detection and averaging is the rigorously optimal procedure for detecting a noise-like signal, and it is virtually optimal for a steady sinusoid. The rigorous optimum in the latter case is well-known and depends on the assumed signal power S; if v is the amplitude of the envelope (output of a linear rectifier), then a non-linear rectifier or other device is to be used to convert v to $\log I_0(v\sqrt{2S}/N)$, and the values of the latter are averaged. $I_0(x)$ is the Bessel function $J_0(x\sqrt{-1})$. This corresponds very nearly to square-law rectification when S/N is small, and linear rectification when S/N is large. Since square-law is little different from linear rectification, as noted in Section 1, it is also little different from the optimal. For a steady dc signal, the optimal procedure uses the average of the algebraic values of the samples (without any rectification); the performance is given by (12a).